A MAP Criterion for Detecting the Number of Speakers at frame level in Model-based Single-Channel Speech Separation

Mowlaee, Pejman; Christensen, Mads Græsbøll; Tan, Zheng-Hua; Jensen, Søren Holdt

Published in:
Asilomar Conference on Signals, Systems and Computers. Conference Record

DOI (link to publication from Publisher):
10.1109/ACSSC.2010.5757617

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

A MAP CRITERION FOR DETECTING THE NUMBER OF SPEAKERS AT FRAME LEVEL IN MODEL-BASED SINGLE-CHANNEL SPEECH SEPARATION

P. Mowlaee[1]*, M. G. Christensen*[2]*, Z. -H. Tan*[1]*, and S. H. Jensen*[1]

[1] Dept. of Electronic Systems, [2] Dept. of Architecture, Design & Media Technology
Aalborg University, Denmark
{pmb,zt,shj}@es.aau.dk mgc@create.aau.dk

ABSTRACT

The problem of detecting the number of speakers for a particular segment occurs in many different speech applications. In single channel speech separation, for example, this information is often used to simplify the separation process, as the signal has to be treated differently depending on the number of speakers. Inspired by the asymptotic *maximum a posteriori* rule proposed for model selection, we pose the problem as a model selection problem. More specifically, we derive a multiple hypotheses test for determining the number of speakers at a frame level in an observed signal based on underlying parametric speaker models, trained a priori. The experimental results indicate that the suggested method improves the quality of the separated signals in a single-channel speech separation scenario at different signal-to-signal ratio levels.

Index Terms— Double-talk detection, single-channel speech separation, multiple-hypothesis test.

1. INTRODUCTION

An open problem in speech processing is the detection of the number of speakers present in a given segment of a signal. A special case of this problem is the classification of speech segments into what is often referred to as single-talk (one speaker), double-talk (speech mixture), and noise-only regions, with the resulting detector commonly referred to as a double-talk detector. Knowledge of such regions is useful since in many speech applications, it is required to process the underlying signals differently depending on the type. In this regard, a detector solving this problem can be effectively used as a pre-processor for improving the performance.

Double-talk detection has been used for a number of applications, two examples being acoustic echo cancellation and single-channel speech separation (SCSS). In acoustic echo cancellation, the double-talk detector is used to freeze the adaptation of an adaptive filter during double-talk regions (when both far-end and near-end speech is present) in order to avoid divergence of the adaptive filter, and, as a consequence, avoid the cancellation of the desired speech signal [1]. However, in SCSS, it is used to classify an observed speech mixture into single-talk, double-talk, and noise only regions, regions that have to be processed differently.

In the context of SCSS, a few separation methods implicitly detect double-talk regions in various contexts, e.g., [2, 3, 4]. In [2], a state-based hypothesis test was proposed in order to determine the reliability of each time-frequency cell in a given noise-corrupted speech signal. It was observed that the method led to a significant improvement in speech recognition performance in presence of other competing speaker signals. Similarly, in [4], a silence state was added to the speaker codebooks in order to deal with frames where only one speaker is active.

A few participants in the speech separation challenge [5], made use of a model-based speaker identification (SID) module, called *Iroquois* [3] to identify speakers existing in the mixture. *Iroquois* works based on excluding silence and mixture segments from its parameter update procedure. Instead, it selects segments where only one speaker is dominated which are known as discriminating features for speaker recognition purpose. This decision-taking helped narrowing down what speakers are present in the mixture, hence, leading to an improvement in speaker recognition performance [3]. This required the calculation of speaker posteriors for different trained models of speakers present in the whole dataset (e.g. 34 speakers in [5]). *Iroquois* used a fixed threshold for calculating the uncertainty in speaker identification, and, as a consequence, could result in errors while determining which frame belongs to single-talk and double-talk regions.

Source-driven approaches, mostly known as computationally auditory scene analysis (CASA) [6], suggest to combine time-frequency segments of the mixed signal that are likely to arise from the same source and then concatenate them into a single stream. As a consequence, CASA-based methods implicitly detect the number of speakers in the speech mixture independently of *a priori* knowledge of any speaker model [6]. However, the methods predominantly use estimated pitch trajectories by applying a multi-pitch estimator. For the masked signal, as a consequence, the overall accuracy for CASA-based method is limited by the accuracy of the multi-pitch estimator.

To solve the problem of detecting the number of speakers in a speech mixture, we take a different approach. We integrate the *maximum a posteriori* (MAP) criterion proposed in [7] into SCSS to solve the model selection problem. We derive multiple hypothesis tests to determine double-talk/single-talk regions in segments of the mixed signal. We present the results of signal classification by applying the proposed method to speech mixtures composed of two speakers at different signal-to-signal ratio (SSR) levels. In addition, to put the idea into perspective, we demonstrate how using the proposed detector will affect the quality of the separated output signals. More specifically, by finding single-talk regions thanks to a double-talk detector, the remaining problem to be solved in SCSS is only to separate the mixture segments. For single-speaker frames, the observed signal is directly re-synthesized according to the corre-

All correspondence should be directed to P. Mowlaee, Dept. of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220, Aalborg, Denmark, email: pmb@es.aau.dk, phone: +45 9940 9888. The work of P. Mowlaee is supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175.

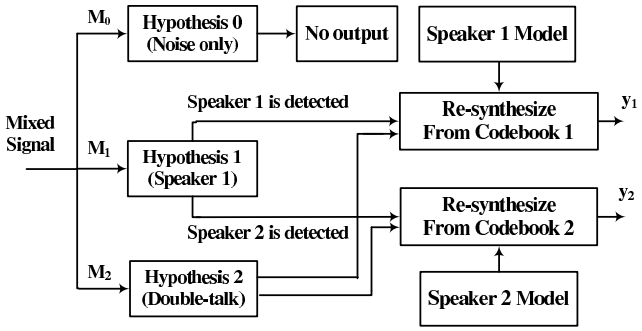

Fig. 1. The schematic block diagram for the proposed method for detecting the number of speakers in mixture and showing how it can be used in the SCSS problem. The decision lies in one of the following three models: $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ showing, noise-only, single-talk, and double-talk classes, respectively. The separated output signals are shown as $\mathbf{y}_1$ and $\mathbf{y}_2$ for speaker one and two, respectively.

sponding speaker models.

The paper is structured as follows: In the next section, we introduce basic notation, definitions and the model-selection problem. In Section 3, we derive multiple-hypothesis rules required for detecting single-talk and double-talk regions in a segment of mixture. In Section 4, we present the experimental results with showing the accuracy of the proposed method. We also present the results showing the improvements achieved by employing the proposed double-talk detector in a SCSS scenario. Section 5 concludes on the work.

2. MODEL SELECTION FOR DETECTING THE NUMBER OF SPEAKERS

We will now proceed to introduce some basic notation and definitions. Consider a mixed signal with N samples $\mathbf{y} \in \mathbb{R}^N$ composed of up to J speaker signals as $\mathbf{y} = \sum_{j=1}^{J} \mathbf{s}(\boldsymbol{\psi}_j) + \mathbf{e}$, where the superscript T represents the matrix transpose, $j \in [1, J]$ the number of signals in the mixed signal, $\mathbf{s}(\boldsymbol{\psi}_j) \in \mathbb{R}^N$ the jth signal characterized by parameter vector $\boldsymbol{\psi}_j$ and $\mathbf{e} \in \mathbb{R}^N$ the noise signal incorporated in the model. For simplicity in the ensuing derivations and simulations, we focus on $J = 2$, that is, a mixture of two speakers and noise. As our signal model, we use sinusoidal modeling as described in [8]. More specifically, we model the jth speaker signal in the mixture as a parametric feature vector $\boldsymbol{\psi}_j$, composed of sinusoidal parameters: amplitude, frequency and phase vectors. We here use $K = 3$ candidate models each denoted by $\mathbf{M}_k$, for describing the mixed signal, $\mathbf{y}$, namely: $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ to indicate noise-only, single-talk, and double-talk, respectively. Each of these models is described by parameter vector $\boldsymbol{\theta}_k$ with L_k sinusoids. A block diagram of the proposed method for detecting the number of speakers in mixture is shown in Fig. 1. The proposed approach addresses the following problem: given the mixed signal, select the model which is the most likely. We consider three models for $\mathbf{y}$ as:

$\mathbf{M}_0$**:** $\mathbf{y} = \mathbf{e}$,
$\mathbf{M}_1$**:** $\mathbf{y} = \mathbf{s}(\boldsymbol{\psi}_j) + \mathbf{e} \quad \text{for} \quad j \in [1, 2]$,
$\mathbf{M}_2$**:** $\mathbf{y} = \mathbf{s}(\boldsymbol{\psi}_1) + \mathbf{s}(\boldsymbol{\psi}_2) + \mathbf{e}$,

where $\mathbf{s}(\boldsymbol{\psi}_1) + \mathbf{s}(\boldsymbol{\psi}_2)$ represents an estimate for the mixed signal, and $\mathbf{s}(\boldsymbol{\psi}_j)$ with $j \in [1, 2]$ indicates the jth signal modeled by the parameter set $\boldsymbol{\psi}_j$.

Following the model selection approach in [7], we adopt a MAP criterion for multiple-hypothesis tests to determine double-talk/single-talk regions in segments of a mixed signal. To this end, we need to evaluate the posterior probabilities of $\mathbf{M}_k$ with $k \in Z_K = \{0, 1, 2\}$. The MAP estimate of the most likely hypothesis is denoted by $\hat{\mathbf{M}}_k$, and is obtained as

$$\hat{\mathbf{M}}_k = \arg \max_{\mathbf{M}_k : k \in Z_K} \left\{ \int_{\boldsymbol{\theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, \mathbf{M}_k) p(\boldsymbol{\theta}_k|\mathbf{M}_k) d\boldsymbol{\theta}_k \right\}. \quad (1)$$

The problem in (1) is a complicated nonlinear maximization problem due to the used models. As proposed in [7], instead of numerical integration for the evaluation of marginal density in (1), we employ the asymptotic MAP criterion, which under certain conditions can be shown to be

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k : k \in Z_K} \left\{ -\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathbf{M}_k) + p_c \right\}, \quad (2)$$

with p_c being the model-dependent penalty of the MAP criterion, $\hat{\boldsymbol{\theta}}_k$ an estimate of $\boldsymbol{\theta}_k$ for the kth model $\mathbf{M}_k$, and $-\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathbf{M}_k)$ the log-likelihood term obtained from an approximation of (1).

3. MULTIPLE-HYPOTHESIS ALGORITHM

The problem is now to determine $-\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathbf{M}_k)$ for each of the three underlying candidate models $\mathbf{M}_k$ with $k \in Z_K = \{0, 1, 2\}$. Here, we use sinusoidal modeling in [8] to model the speaker signals in the mixture. Let $\mathbf{s}_i(\hat{\boldsymbol{\psi}}_j)$ be the jth speaker signal with $j \in [1, 2]$ for the ith frequency band modeled by the parametric vector $\hat{\boldsymbol{\psi}}_j$. Here we assume that the signal modeling error, $\mathbf{e}$ has a Gaussian distribution and the modeling error subband signal, $\mathbf{e}_i$ is white in each ith frequency band. Then from the subband decomposition and the independece assumption for all frequency bands, assuming that $\mathbf{e}_i$ is independent from one band to another, one can show that the likelihood function for all bands for each class $\mathbf{M}_k$ is given by

$$\begin{aligned} p(\mathbf{e}|\sigma^2) &= \prod_{i=1}^{Q} p(\mathbf{e}_i|\sigma_i^2) \\ &= \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{i=1}^{Q} \sigma_i} \exp\left(-\frac{1}{2} \sum_{i=1}^{Q} \frac{\mathbf{e}_i^T \mathbf{e}_i}{\sigma_i^2} \right), \end{aligned} \quad (3)$$

where Q is the total number of frequency bands, σ_i denotes the variance due to the modeling error signal in the ith band, $\mathbf{e}_i$.

For single speaker class, $\mathbf{M}_1$, the modeling error at the ith frequency band, is given by $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{s}_i(\hat{\boldsymbol{\psi}}_j)$. For the mixed class, $\mathbf{M}_2$, let us define the estimated error as $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{s}_i(\hat{\boldsymbol{\psi}}_1) - \mathbf{s}_i(\hat{\boldsymbol{\psi}}_2)$ as the noise estimated for the ith frequency band as a colored noise not fitted by $\mathbf{M}_2$. The MAP criterion [7] for sinusoids composed of unknown amplitudes and frequencies reduces to

$$\hat{\mathbf{M}}_k = \arg \min_{\mathbf{M}_k \in Z_K} \left\{ \frac{N}{2} \sum_{i=1}^{Q} \ln \hat{\sigma}_i^2 + \frac{5L_k}{2} \ln N \right\}. \quad (4)$$

where we define $\hat{\sigma}_i^2 = \frac{1}{N} \hat{\mathbf{e}}_i^T \hat{\mathbf{e}}_i$ as the estimated variance for the ith frequency band and we remind the reader that L_k is the number of sinusoids. In the mixture class $\mathbf{M}_2$, we require a mixture estimate to replace $\mathbf{s}(\hat{\boldsymbol{\psi}}_1) + \mathbf{s}(\hat{\boldsymbol{\psi}}_2)$ in order to find the best pair of $\{\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2\}$ from the speaker models of the underlying speakers. Here, we use the minimum mean square error (MMSE) estimator for the mixture magnitude spectrum in [9], in order to find the the joint best states in the speaker models which when combined best describe the magnitude spectrum for the observed mixture, $\mathbf{y}$.

Table 1. Speaker labels used for training the gender-dependent models for male and female speakers.

Male	3	5	6	9	10	12	13	14	17	19
Female	4	7	8	11	15	16	21	22	23	24

We include the noise model, $\mathbf{M}_0$ as one of the examined models by setting $\mathbf{y} = \hat{\mathbf{e}}$ and setting the number of sinusoids equal to zero ($L_k = 0$). The estimated noise variance is given by $\hat{\sigma}_i^2 = \frac{1}{N}\mathbf{y}_i^T\mathbf{y}_i$.

Finally, using the estimated value for σ_i depending on each possible class of $\mathbf{M}_k$ with $k \in Z_K = \{0, 1, 2\}$, the best model, as a result, is the one which yields high log-likelihood and low model order, which is achieved in (4). The proposed method for detecting the number of speakers in the speech mixture can be summarized in the following three steps:

(1) Find the variance of noise, $\hat{\sigma}_i$ at each ith band.
(2) Compute the MAP criterion for each class: $\{\mathbf{M}_0, \mathbf{M}_1, \mathbf{M}_2\}$.
(3) Select the model with largest log-likelihood.

4. SIMULATION RESULTS

4.1. System Setup and Database

To evaluate the proposed approach, we used the database in [5] with a sampling rate of 8 kHz. The speaker models are obtained by the split-VQ (vector quantization) [8] composed of sinusoidal amplitude and frequencies trained based on 10 minutes of speech signals for each speaker. For training the speaker models we used 2048 codevectors for amplitude and 8 codevectors for frequency part. Throughout the experiments, a Hamming window of length 32 ms with frame-shift equal to 8 ms was used to segment the speech files both in the training and test phase. As our test data, we used the mixture of target and masker speakers in the test setup of [5] mixed at six SSR levels of {-9,-6,-3,0,3,6} dB. To relax the speaker-dependent assumption, we used gender-dependent models and we trained a male speaker model using utterance from ten speakers and a female speaker model trained on ten female speakers. The speaker labels used for training our gender-dependent models are shown in Table 1.

4.2. Experiment 1: Detection Accuracy

Figure 2, shows the clean signal (prior to mixing) for speaker one and two together with their mixture. In Fig. 2, the detection results of the number of speakers in speech mixture are shown for gender-dependent scenario. The hypotheses for single-talk and double-talk regions are also shown as ground truth. It is observed that, the double-talk detector effectively finds the regions of the non-speech and mixture segments and determines at each frame that which speaker(s), if any, are active. Comparing with the ground-truth, it is observed that the accuracy of the proposed double-talk detector is high. In our experiments, the models $\mathbf{M}_k$ with $k \in Z_K$ are considered as either speaker-dependent or gender-dependent. It is important to note that, in the speaker-dependent scenario, the proposed method solves a four class problem, namely noise, speaker one, speaker two, and mixture classes. However, using gender-dependent speaker models, the proposed double-talk detector solves a three-class problem for same gender or same talker scenario, since the estimated error signal, given by single-talk classes, will be the same.

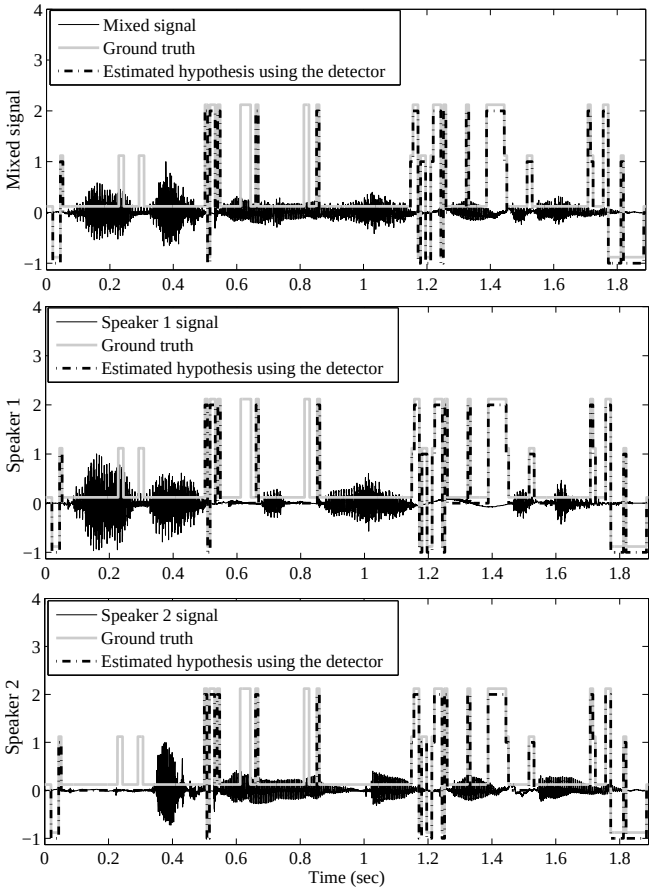

Fig. 2. Showing the performance for detecting the number of speakers in a mixture of a male and a female speaker mixed at 3 dB SSR. The mixed signal is composed of a male (speaker 12) uttering *"Lay white with e 8 again"* with female (speaker 11) uttering *"Set green with v 3 soon"*. Decisions are -1 for no speech, 1 for speaker one, 2 for speaker two and 0 for mixed signal regions.

4.3. Experiment 2: Speech Separation

In another experiment, we aim to study the effectiveness of employing the proposed double-talk detector in a SCSS system. More specifically, as a proof of concept, we report the signal quality of the separated signals obtained by using a model-based separation system with and without double-talk detector proposed in this work. Figure 3 shows the perceptual evaluation of speech quality (PESQ) [10] scores averaged over 50 mixtures The results are reported for both speaker-dependent and gender-dependent scenarios. From the PESQ curves shown in Fig. 3, it is observed that integrating double-talk detector into a model-based SCSS improves the speech quality of the re-synthesized signals. It is also observed that the PESQ scores obtained in the gender-dependent scenario were slightly lower than those obtained in speaker-dependent scenario. However, as the SSR increases the performance of gender-dependent scenario aysmptotates to the one offered by speaker-dependent scenario. From informal listening test, it was observed that, the improvement obtained by employing the proposed detector is noticeable.

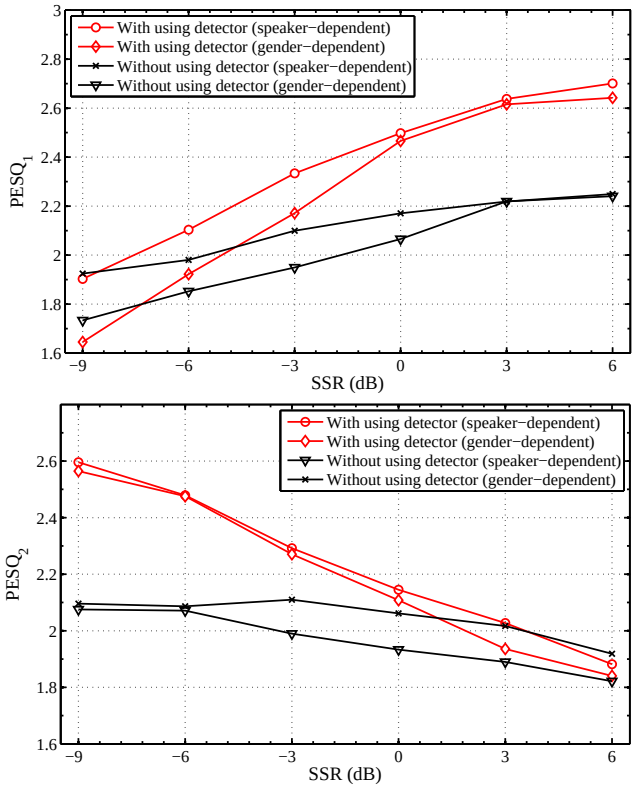

Fig. 3. Showing the PESQ scores obtained for speech separation in speaker-dependent and gender-dependent scenarios for two cases: with and without using the proposed method for detecting the number of speakers in a given speech mixture: (top panel) the PESQ scores for the first speaker and (bottom panel) for the second speaker in terms of the SSR level in decibels.

5. CONCLUSION

To conclude on our work, we have presented a solution to detecting the number of speakers in an observed segment of mixed speech signal. To solve the problem, we applied the MAP criterion already proposed for model selection and derived the multiple-hypothesis test algorithm to determine double-talk/single-talk regions for a particular segment in a given mixed signal in SCSS framework. We showed that, such information can be used to narrow down the separation problem only for mixed frames. Experiments showed that the proposed method successfully determines the single-talk and double-talk regions in both speaker-dependent and gender-dependent scenarios. The proposed detector approach also led to improvement in the signal quality of the separated signals compared to the scenario where no detector was used.

6. REFERENCES

[1] J. Benesty, D. Morgan, and J. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 168 –172, Mar. 2000.

[2] S. Srinivasan and D. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, March 2005, pp. 89–92.

[3] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[4] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 67 – 76, Jan. 2010.

[5] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, Jan. 2010.

[6] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 77 – 93, Jan. 2010.

[7] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[8] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, Aug. 2008.

[9] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Sinusoidal masks for single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2010, pp. 4262–4266.

[10] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 749–752, Aug. 2001.